A Big Data Framework to Address Building Sum Insured Misestimation

Roberts, Callum; Gepp, Adrian; Todd, James

Big Data Research

DOI:
10.1016/j.bdr.2023.100396

Published: 28/08/2023

Peer reviewed version

A Big Data Framework to Address Building Sum Insured Misestimation

Abstract

In the insurance industry, the accumulation of complex problems and volume of data creates a large scope for actuaries to apply big data techniques to investigate and provide unique solutions for millions of policyholders. With much of the actuarial focus on traditional problems like price optimisation or improving claims management, there is an opportunity to tackle other known product inefficiencies with a data-driven approach. The purpose of this paper is to build a framework that exploits big data technologies to measure and explain Australian policyholder Sum Insured Misestimation (SIM). Big data clustering and dimension reduction techniques are leveraged to measure SIM for a national home insurance portfolio. We then design predictive and prescriptive models to explore the relationship between socioeconomic and demographic factors with SIM. Real-world results from a national home insurance portfolio provide actionable business insight on SIM and facilitate solutions for stakeholders, being government and insurers.

Keywords: clustering large applications; home insurance; sum insured estimation; underinsurance; big data

1. Introduction

In the technology age, insurers are collecting and outsourcing a wealth of data at an increasing rate. They use this information, ranging from geospatial to telematics data, to improve various analytical functions of the business, ultimately increasing profitability and providing the customer with a better product and experience. Much of this work is focused on core business processes like pricing and valuation, and this leaves a large scope in the industry to explore the application of big data techniques in other areas.

A recent report by the Australian Competition and Consumer Commission (ACCC) provided a recommendation that estimating the sum insured (SI) is one area where insurers could, and should, provide better guidance to consumers to lessen the risk of underinsurance [1]. For some time, the industry has recognised home underinsurance as a severe problem. In Australia, where frequent natural catastrophe activity can cause many concentrated total losses, underinsurance is rampant [2]. There have been substantial steps taken by the industry to address this problem. A recent paper by the Actuaries Institute on Property Insurance Affordability developed an affordability measurement and provided non-pooling methods to increase insurance affordability [3]. Some insurers provide an increase in nominated sum insured if that value is insufficient to cover costs, and other larger insurers will elect to cover any loss arising[1]. These approaches are positioned from a qualitative standpoint, and it is difficult to validate their effectiveness.

Big data analytics are a clear avenue to investigate the portfolio-wide problem of Sum Insured Misestimation (SIM). The desirable outcome in addressing SIM is that insurers will be able to better

[1] For example, “Complete Replacement Cover” offered by AAMI insurance as per https://www.aami.com.au/aami/documents/personal/home/aami-home-building-insurance-pds.pdf

quantify the likelihood and the amount by which consumers misestimate their SI to ensure that customers at claim time will not be left under compensated. Additionally, statistical techniques can be used to explain the driving factors of SIM, which aids governments in devising pooling methods, subsidies, and risk mitigation measures. Appreciating the social consequences of algorithm prediction in insurance discussed in [4], this research provides an application of big data analytics that leads to improved social outcomes. The motivation to measure SIM will remain for long as insurers are unable to comfortably rely on customers or third-party estimates of SI.

This paper presents a big data framework for analysing, predicting, and explaining SIM for home insurance. A two-pronged approach is taken with parts of the framework, with sampling used to maximise exploration of the dataset, and big data techniques to scale results for practical implementation. Using big data analytics, the framework aims to address the crucial part of the SIM problem, in that we do not directly observe the true SI should be for any customer. The framework's components were also influenced by the analytical big data frameworks presented in [5-6] where they use predictive and prescriptive analyses to facilitate social and business action. The key parts of the framework are explained below:

- Clustering Analysis: an unsupervised learning algorithm designed for big data, Clustering Large Applications (CLARA), is used to address the key aspect of omitting bias in customer selected SI values by grouping similar insured buildings using building features only. Additionally, this algorithm helps insurers reduce the dimensionality and noise in big datasets.
- Predictive Analysis: we propose a measure for SIM which is calculated analytically from the clustering results. Statistical prediction models are then used to explain and identify policyholders likely to misestimate their SI. For exploratory purposes, we then use SHapley Additive exPlanations (SHAP) values to produce waterfall charts, importance plots and explain complex relationships driving SIM.
- Prescriptive Analysis: provides context and packages the outcomes from predictive analysis for the benefit of stakeholders. SIM is an important issue for insurers and government, and we note future research would look to consolidate the prescriptive analysis findings from the application of this framework with other home insurance portfolios.

The remainder of this paper is structured as follows: a summary of relevant big data analytics applications in insurance literature is provided in Section 2. An overview of the SIM problem in literature is also included in Section 2. Section 3 introduces the datasets used in this work. The big data framework developed for the analysis, prediction and explanation of property SIM is discussed in Section 3. Section 4 includes thorough detail on the validation and experimental results of the framework. The work is concluded in Section 5, with recommendations for future analysis.

2. Literature review

There is limited academic literature in the SIM space. In total, 4 papers were returned from using search terms "underinsurance", "under insured" in the Web of Science database. We excluded any papers that referred to insurance products other than retail Home Insurance. We also conducted a review on papers that used big data clustering techniques with applications in insurance, to cater for the big data analytics piece of this work.

Booth and Tranter [7] provided insights into associations and unfolding effects of house and contents underinsurance in Australia. They reproduced underinsurance along socioeconomic and geographic lines, with those of lower socioeconomic status or living in cities more likely to be underinsured. They also identified factors that explain house insurance uptake, such as age, income, education, and marital status. In a case study on insurance vulnerability in rural Australia, Whittaker et al. [2] found that many

residents and landlords were substantially underinsured for damage to livelihood assets such as farm fences, livestock, and sheds. The paper proposed that cultural, economic, political, and social factors prevent people from attaining an adequate level of cover. Hope [8] found that the most vulnerable policyholders to underinsuring were those in high-risk categories for natural phenomena. Grislain-Letrémy [9] hypothesised that the margin in the insurance market was due to uninsurable housing and the anticipation of financial assistance. The focus of existing SIM literature is on the risk factors correlated with underinsurance and understanding the impacts of underinsurance on populations. There were no papers found that considered over-insurance or made an attempt to use quantitative information to estimate SIM.

From a broader review of literature around clustering within insurance, and to the best of the authors' knowledge, no prior research has used clustering algorithms, or any big data techniques, for the SIM problem. We found that clustering applications in insurance involved rate making, policyholder behavioural analysis, risk characteristic analysis and geographical risk analysis. Additionally, this research predominantly used the k-means algorithm, where it was applied to small scale problems. The algorithms and sample sizes used in the relevant research are summarised in **Table 1** and **Table 2**. Yeo and Smith [10] used k-means clustering to separate automobile insurance policyholders based on risk characteristics. The groupings were used to analyse customers' overall price sensitivity to premiums through modelling retention rates. Samizadeh and Mehregan [11] used k-means clustering to cluster policyholders in a life insurance company to predict customer behaviour based on payment type. Kašćelan et al. [12] used k-means clustering to identify homogenous groups of car insurance policyholders based on their risk characteristics to improve risk premium calculations. K-means clustering was also used by Williams and Huang [13] for identification of policy owners with large claim sizes.

Several papers considered more sophisticated clustering techniques. Khalili-Damghani et al. [14] used a two-stage clustering classification model to recommend suitable insurance coverages to customers. Dehghanpour and Rezvani [15] apply Similarity Based Agglomerative Clustering (SBAC) to find different segments and associated attributes of insurance fraud. Carfora et al. [16] explored the use of K-means and other K-means derived clustering algorithms (Cobweb, FarthestFirst) to identify driver behaviour parameters that can be incorporated into methodologies for automobile insurance coverage. Devale et al., [17] used a SBAC algorithm for segmenting and attracting new customers.

Zhuang et al. [18] used Partitioning Around Medoids (PAM) and SBAC in a study on insurance customer segmentation and explored the algorithm's functionality with mixed-data type similarity measures. Yao [19] compared a wide range of partitional, hierarchical and density-based clustering, including CLARA, in their application to insurance ratemaking. Kharel et al. [20] used a Poisson mixture methodology to develop and model clusters of natural hazard phenomena in general insurance.

The gap identified here is that firstly, SIM research has not been approached from a purely quantitative angle, and there has been no consideration for big data or associated techniques. These techniques, specifically clustering techniques have been useful in other areas of insurance for similar tasks, and are a natural choice for this problem given the unsupervised nature of SIM. Furthermore, there was only one example of a large application algorithm, CLARA, in insurance. This study also deeply contributes to SIM research – current SIM literature is limited to specific case studies that are not applicable to the wider population such as rural buildings [7] or using anecdotal evidence as a proxy for SIM) [8]. Taking a quantitative angle with a national insurance portfolio, provides a wealth of statistical bases from which to derive information about SIM – which aims to result in more reliable estimates of SIM that are applicable to a wider population of buildings than investigated in current literature. Our big data analytics framework for SIM aims to address the lack of quantitative research for SIM and contribute to the wider initiative of big data clustering techniques in insurance.

	Partition						Hierarchical						Density-Based		
	K-Means	K-Medoids (PAM)	Improved K-Medoids	CLARA	CLARANS	Poisson-mixture	AGNES	DIANA	BIRTH	CURE	CHAMALEON	SBAC	DBSCAN	OPTICS	DENCLUE
[15]	✓											✓			
[17]												✓			
[12]	✓														
[14]	✓														
[20]						✓									
[11]	✓														
[19]	✓	✓		✓	✓		✓	✓	✓	✓	✓		✓	✓	✓
[10]	✓														
[18]		✓	✓									✓			
[13]	✓														
[16]	✓														
TOTAL (Clustering Algorithm)	**8**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**3**	**1**	**1**	**1**
TOTAL (Clustering Method)	**14**						**8**						**3**		

Table 1 Summary of clustering algorithms used in papers. The algorithm definitions are not strict as to reduce the number of total algorithms presented in the table. For example, K-means includes algorithms that use some element of K-means in the wider algorithm such as Cobweb and FarthestFirst in [16].

Sample Size	Papers	Relative Frequency	Cumulative Frequency
< 1000	3	30.0%	30.0%
1001 to 10,000	3	30.0%	60.0%
10,001 to 100,000	3	10.0%	70.0%
100,000+	1	30.0%	100.0%
Total	10	100.0%	100.0%

Table 2 Samples sizes of the papers reviewed that applied clustering algorithms. For comparison, this study uses a dataset of approximately 400,000 observations.

3. Methodology

In this work, we present a framework to assess the misestimation of a policyholder's building SI value. The current home insurance portfolio of a major Australian insurer was selected as the case study. The framework for SIM is shown in **Fig. 1** and constitutes of three main stages: clustering, predictive, and prescriptive analysis.

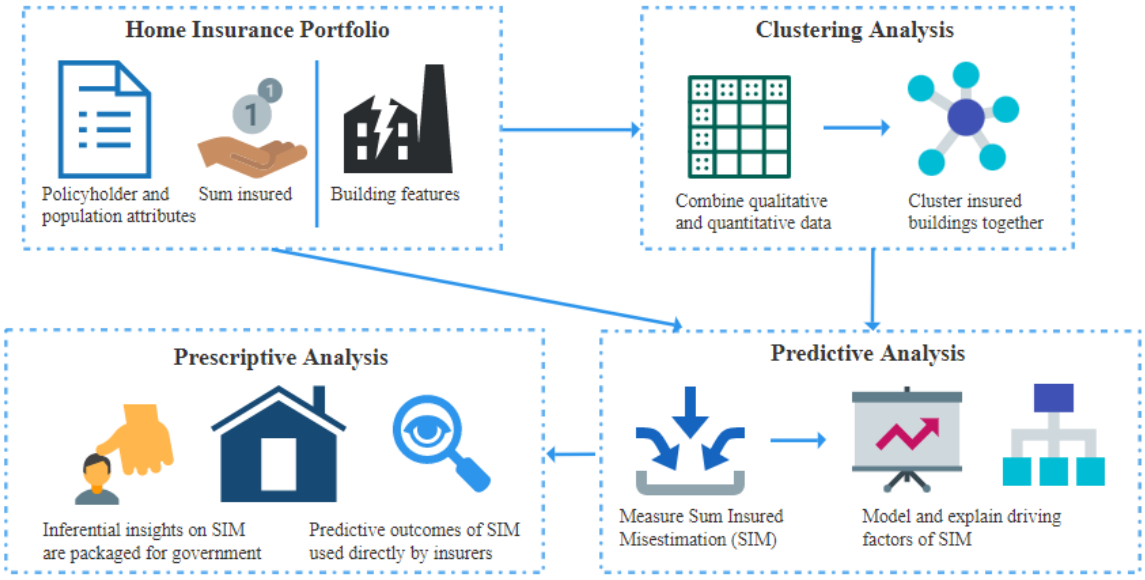

Fig. 1. The big data framework for analysing, predicting, and explaining SIM for property insurance. The dataset used should be a reasonable approximation of the population, i.e. a national portfolio. The results from the Clustering Analysis along with the policyholder and sum insured data feed into the Predictive Analysis. Using the results from the Predictive Analysis, Prescriptive Analysis is provided for relevant stakeholders. For insurers, Prescriptive Analysis can lead to solutions requiring real-time implementation and these can be facilitated by storing predictive results on cloud servers.

3.1 Data

The data set sourced from an Australian insurer consists of approximately 400,000 risks[2] (sold policies) collected in May 2021. In the quote process the policyholder will manually enter their personal information and a SI value for their insured building within the appropriate underwritings bounds. Via a unique identifier, individual risk, geographic, socioeconomic, and demographic information are attached to the risks. The information required at the various analysis stages is provided in **Table 3** and **Table 4** along with a description of each variable.

For the clustering analysis stage of the framework, we derive a list of rebuilding variables that should influence the true SI value. These variables are further categorised by common attributes such as height, size, and building quality. For predictive and prescriptive analysis, relevant socioeconomic and demographic features are used. The Sum Insured Cover is also listed, which indicates whether a policyholder has opted for cover in the event they are underinsured. It can be seen from **Table 4** that several features are captured at the population level and thus more importance is placed on the features at the policyholder level.

Fig. 2 presents the mean SI for the features that provided the greatest differentiation in SI to demonstrate the suitability of the features listed for use in the clustering algorithm. **Fig. 3** displays the distribution of SI values for the dataset.

[2] Exact number of observations not provided due to commercial confidentiality.

ID	Variables	Type
Building Height		
1	Eave Height	Continuous (in metres)
2	Roof Height	Continuous (in metres)
3	Slope of Land	Continuous (in metres)
Building Size		
4	Number of Bedrooms	Categorical, (1-10 and Unknown)
5	Number of Bathrooms	Categorical, (1-10 and Unknown)
6	Building Area	Continuous (in square metres)
Building Quality/Type		
7	Wall Type	Categorical (11 levels*)
8	Roof Type	Categorical (5 levels*)
9	Year Built	Continuous (1800 to 2021)
10	Building Type	Categorical (10 levels*)
11	Security Systems	Categorical (6 levels*)
12	Swimming Pool	Categorical (Yes & No)

Table 3 Description of building variables used in descriptive and clustering analysis. Details are omitted for variables indicated by a "*" due to commercial sensitivity.

Var. No.	Variables	Type^
Policyholder Level		
1	Age	Continuous (in years)
2	Gender	Categorical (Female, Male and Unknown)
3	Tenure at address	Continuous
4	Retired	Categorical (Yes, No and Unknown)
5	Median Weekly Rent	Continuous (in AUD)
6	Sum Insured Gap Cover	Categorical (Yes, No)
Population Level		
7	Weekly family income	Percentile (at Postcode)
8	Socio-economic factor 1	Percentile (at SA1 level)
9	Socio-economic factor 2	Percentile (at SA1 level)
10	Socio-economic factor 3	Percentile (at SA1 level)
11	Socio-economic factor 4	Percentile (at SA1 level)
12	Socio-economic factor 5	Percentile (at SA1 level)
13	Population density	Population rate per sqm (at SA1 level)
14	Youth population	Youth population out of total population (at SA1 level)
15	Crime factor 1	Rate per 100k (at Police District)
16	Crime factor 2	Rate per 100k (at Police District)
17	Crime factor 3	Percentile of rate (at Police District)
18	Crime factor 4	Percentile of rate (at Police District)
19	Drug factor 1	Decile of rate (at Police District)
20	Drug factor 2	Decile of rate (at Police District)
21	Language proficiency (English, other)	Percentile (at SA1 level)
22	Highschool education achieved	Percentile (at SA1 level)
23	Education type 1	Percentile (at SA1 level)
24	Education type 2	Percentile (at SA1 level)
25	Marital status	Percentile (at SA1 level)
26	Unpaid work/total care for children	Percentile (at SA1 level)
27	Number of motor vehicles owned	Percentile (at SA1 level)
28	Number of residents	Percentile (at SA1 level)
29	Mortgage repayment	Percentile (at SA1 level)
30	Rent and tenure type	Percentile (at SA1 level)
31	Socio-Economic Indexes for Areas Index (SEIFA)	Ranking 1-10 (per Postcode)
32	Index of Relative Socio-economic Advantage and Disadvantage (IRSAD)	Ranking 1-10 (per Postcode)
33	Index of Education and Occupation (IEO)	Ranking 1-10 (per Postcode)

Table 4 Description of socioeconomic and demographic variables used in prescriptive analysis. Some factor names are generalised due to commercial sensitivity, such as “Crime factor X” and “Socio-economic factor X”. Definitions of population groupings that can be provided are:

- Postcode: Geographical postal areas, there are approximately 2,600 postcodes in Australia.
- SA1: Statistical area 1 defined by the Australian Bureau of Statistics (ABS) are designed to maximise the geographic detail available for Census of Population and Housing data. There are approximately 57,500 SA1 levels in Australia.
- Police District: Geographical areas pertaining to police districts, maintained by each Australia state – for example, some states use Local Government Areas (LGAs).

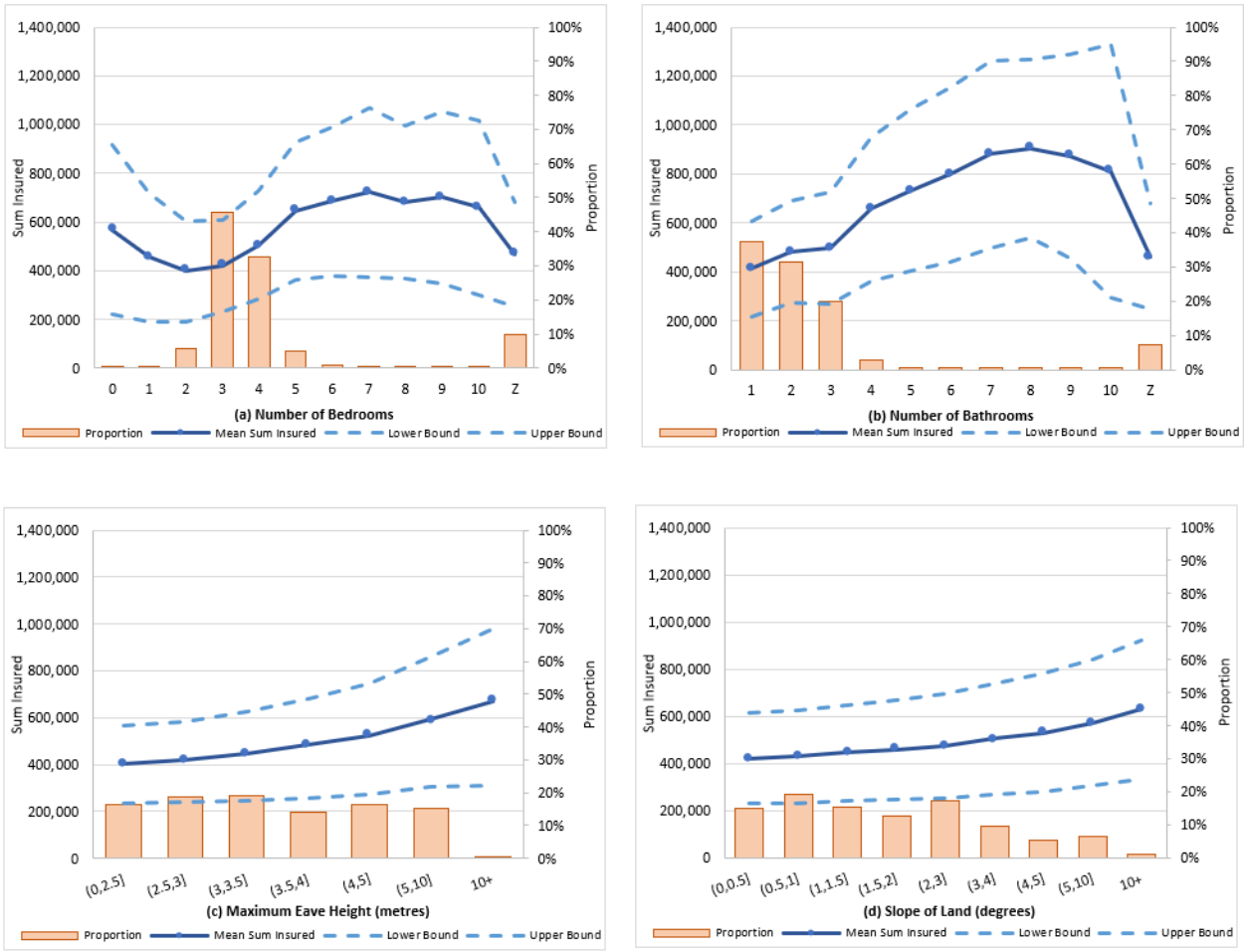

Fig. 2. Using the univariate analysis method, the relation between the SI value is presented for the four attributes that differentiated the SI value the most. The lower and upper bounds reflect one standard deviation of the SI value. The sample figures demonstrate intuitive results, with a higher mean SI for buildings with a larger number of bedrooms and bathrooms, taller (eave height), and more complex structures (slope of land). These relationships strengthen the argument to use clustering analysis to create an unsupervised representation.

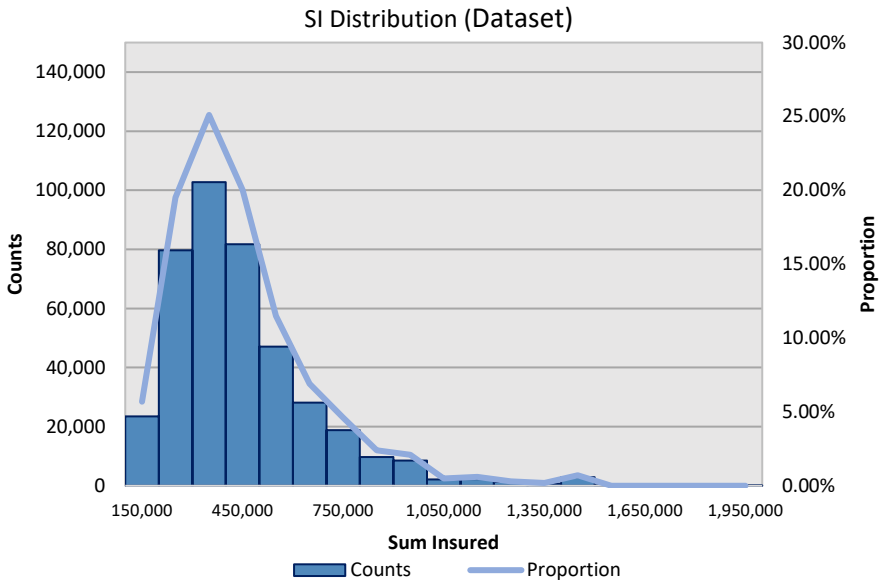

Fig. 3. Distribution of SI value for the dataset.

3.2 Clustering analysis

In this stage, clustering, an unsupervised learning technique, is used to group similar buildings based on their building features. The key aspect of this algorithm is that the building features are selected a-priori, and the policyholder's chosen SI value is removed from the analysis to omit bias. By design, the clusters outputted from the algorithm should represent a specific type of building, with which we can attribute a specific SI value.

3.2.1 Data preparation

Insurance data will usually contain categorial and numerical attributes, so the first constraint that unsupervised learning techniques face is how distance measures can efficiently deal with mixed data types. Several mixed data-type distance measures exist, such as the Goodall measures, but the required CPU power makes implementation of these measures impractical [21]. An alternative approach is to use Factored Analysis for Mixed Data (FAMD) – the mixed data variation of Principal Components Analysis (PCA) – to transform the dataset into squared loadings [22].

FAMD is a computationally feasible and practical solution that allows use of continuous distance measures with mixed data types. FAMD allows for inclusion of all attributes from the original data but in a lower-dimensional representation that is more tractable for clustering techniques. The lower dimensions result in a non-linear decrease in computational time of the distance and clustering calculations. In this work, FAMD was computed on the full dataset, and Manhattan distance was selected for the dissimilarity matrix. FAMD reduced the dataset from 63 dimensions (including one-hot-encoding for factor attributes) to 36, resulting in a 43% reduction in dimension with 70% of variance explained.

3.2.2 Clustering techniques

The aim of the clustering algorithm is to create clusters that accurately represent a specific type of building, such that we can attach a single SI value to all buildings within a cluster. Given the lack of true labels in unsupervised problems, the selection criteria for the clustering technique are influenced more heavily by the constraints of algorithms rather than performance measures.

The crucial constraint in this work is the computational complexity of running the algorithm, and the memory requirements inherent to distance-based clustering, due to the size and breadth of the dataset (n=~400,000). A secondary constraint is the nature of insurance data which contains significant outliers. To account for these big data challenges, we use Partitioning Around the Medoids (PAM) via Clustering Large Applications (CLARA) implementation, referred to jointly as CLARA-PAM[3]. CLARA-PAM is a specifically designed algorithm for large applications and has seen huge improvements in the efficiency of runtime for large datasets [23]. PAM also effectively deals with outliers, when compared with other efficient partitioning approaches such as k-means, by using the median value of a cluster (medoid) as opposed to the mean value (centroid) as the cluster centre to measure distances from [24]. For completeness, we assessed the efficacy of other algorithms on a sample of the dataset to understand if it would be beneficial for future research to consider scaling the application of other algorithms. Partitioning and hierarchical methods were the main focus due to the smaller amount of parameter selection and because they are generally more efficient than density-based methods. Three algorithms other than CLARA-PAM were compared: k-means, AGNES, and H-Clust.

3.2.3 Validation

The final part of clustering analysis involves the validation of the algorithm. This involves the selection of k clusters for CLARA-PAM using relative indexes and performing sensibility checks on the cluster output

[3] To compute the algorithm we used the *cluster* package available on R, a free open-source programming language. Results were collected on a computer with i7 10700K CPU and 64.0 GB of RAM. RAM of at least 16.0 GB was required to compute the CLARA-PAM algorithm. The use of cloud computing resources could be used to further improve the efficiency on large datasets.

using external indexes. We are concerned with finding groups of insured buildings that can reasonably represent a specific house rebuild, and thus indexes that measure intra-cluster similarity, but also incorporate inter-cluster dissimilarity are most suitable. Three relative validation indexes that satisfy these criteria were selected: Silhouette index, Davies-Bouldin index and Dunn index. These indexes were used to determine the number of k clusters in the CLARA-PAM algorithm through a majority vote.

3.3 Predictive analysis

In the predictive analysis a measure of SIM is first devised from the information provided by the clustering analysis; predictive models are then used to explain SIM. First, we take the clusters that each represent a specific type of building and analyse the associated SI values selected by the policyholders. This leads to formulating a proxy measure of SIM. Given we have the response variable SIM, we can then use predictive models as means to investigate relationships that may exist between SIM and the socio-economic and demographic factors. As a by-product of using a national insurance portfolio, the predictive component also encompasses inferential goals.

3.3.1 SIM measure

To formulate the SIM measure, we investigated taking the difference between the policyholders selected SI and some view of each cluster's true SI. In managing the limitation of the true SI value not being known, a level of judgement is used. Such judgements can include focusing on capturing SIM at the extremes, cross-referencing results with related studies [2,7,8,9], and understanding outcomes with actuarial knowledge of SIM [3]. In arriving at a measure for the true SI for each cluster, our main consideration was that outliers are expected to be the buildings that are most over or underinsured, and thus their values should be down-weighted in the true SI calculation. To account for this, we used the Huber-M estimator with a winsorize factor of 1.5, which down-weights non-Gaussian tails of the SI distribution by 1.5 times the standard deviation. Depending on the proportion of over and underinsured buildings within a given cluster, using the Huber-M estimator as the true SI will result in a larger SIM for outlier values, thus capturing SIM at the extremes. We also considered our actuarial knowledge, that each cluster can have differing levels of over and under SIM, which skews common metrics such as the mean – the Huber-M mean better reflects the true SI value by winsorizing the potentially over and underinsured observations. Taking the difference between the policyholders selected SI and the Huber-M estimator for their respective cluster, we arrive at a value of SIM for each policyholder's insured building.

3.3.2 Modelling SIM

Once a measure of SIM is established, predictive analysis serves to identify and predict the SIM for customers in the insurance portfolio, additionally we can obtain inferential insights about the portfolio. In this study, we investigate the relationship between SIM (the response variable) and socioeconomic/demographic features (explanatory variables). For exploratory purposes, the gradient boosting technique, XGBoost, was used to model the relationships. SHAP values are computed to illustrate the relationships, and furthermore we discuss the relevance of each SHAP visualisation. Additionally, the relationships observed between explanatory variables and SIM can be used as a basis to validate the framework and the SIM measure. For example, we can see if the effect that underinsurance is more prevalent in catastrophe prone areas [2] is observed in the predictive models. In this work, we validated our framework with the relationships observed between the SIM measure and wealth, income, and other related economic factors.

3.4 Framework scope and limitations

In practice, Australian insurers do not capture whether a building is underinsured and over-insured; if there data were captured, then it could be used for a supervised learning task. However, insurers are only able to capture this information, for the very small number of buildings that claim a total loss – where still, the cost of a total loss is usually deemed to be the nominated SI, even in cases where the building could have been over or underinsured. Given that securing a reliable dataset suitable for

supervised learning is not feasible, the unsupervised learning approach presented in this paper is proposed as a suitable quantitative way to address the SIM problem.

The factors that lead to SIM are noted in [2,7,8,9] as stemming from demographic, socio-economic, and geographic factors; we have accounted for all of these (as listed in **Table 3**). Based on the case study findings, using our framework should also tend to more robust estimates of SIM relationships due to the size and coverage of a national insurance dataset.

It is relevant to note that there are actuarial considerations that the framework does not capture. The most notable is the demand surge phenomenon for rebuild materials and labor, which inflates the cost of a rebuild and hence the sum insured covered under the policy. A demand surge is experienced when there are a number of total losses concentrated in the same geographical region, usually the result of a catastrophe event such as a severe weather event. Policy terms and conditions generally make allowances for these types of situations, but it still remains a factor in evaluating potential underinsurance. Other less material considerations are factors like inflation and macroeconomic trends for rebuild costs and materials.

4. Experimental results and discussion

4.1 Clustering analysis

We set the sampling amount to 10,000 observations for the CLARA-PAM algorithm. We found that the average distance from medoids for the fixed observations in Table A.1 in the appendix did not materially change with increases in sample size past 10,000 and so did not warrant the increased computational requirements. Smaller sample sizes were found not to allow for the formation of clusters large enough to provide reasonable metrics of SIM. In practice, if the computing power and time is available for a larger sample, or even the full population such that PAM can be used directly (as opposed to the big data implementation CLARA-PAM), then this is encouraged. The extent to which the sample size of 10,000 is generalisable to other studies depends on the coverage of the insurance dataset in terms of homogeneity of risks – for example, a small sample size may suffice when dealing with a state-wide insurance portfolio as more homogeneity is expected with fewer types of buildings (and thus, fewer clusters).

The parameterisation of k in CLARA-PAM is presented in appendix **Fig. B.1**, which shows that diminishing returns are received after ~100 clusters across all validation indexes. A majority vote is taken across indexes with the number of clusters set at 160 to parameterise CLARA-PAM. **Fig. 4** presents the SI distributions for each of the 160 clusters, ordered by the mean SI. From **Fig. 4**, we observe that there is (i) reasonable intra-cluster mean SI variability (ii) high inter-cluster SI variability, especially in the tails of the distributions (iii) and some extreme outliers for most clusters. Observation (i) confirms the findings from the univariate analysis in **Fig. 2**, where (ii) and (iii) suggest that there is a material level of SIM.

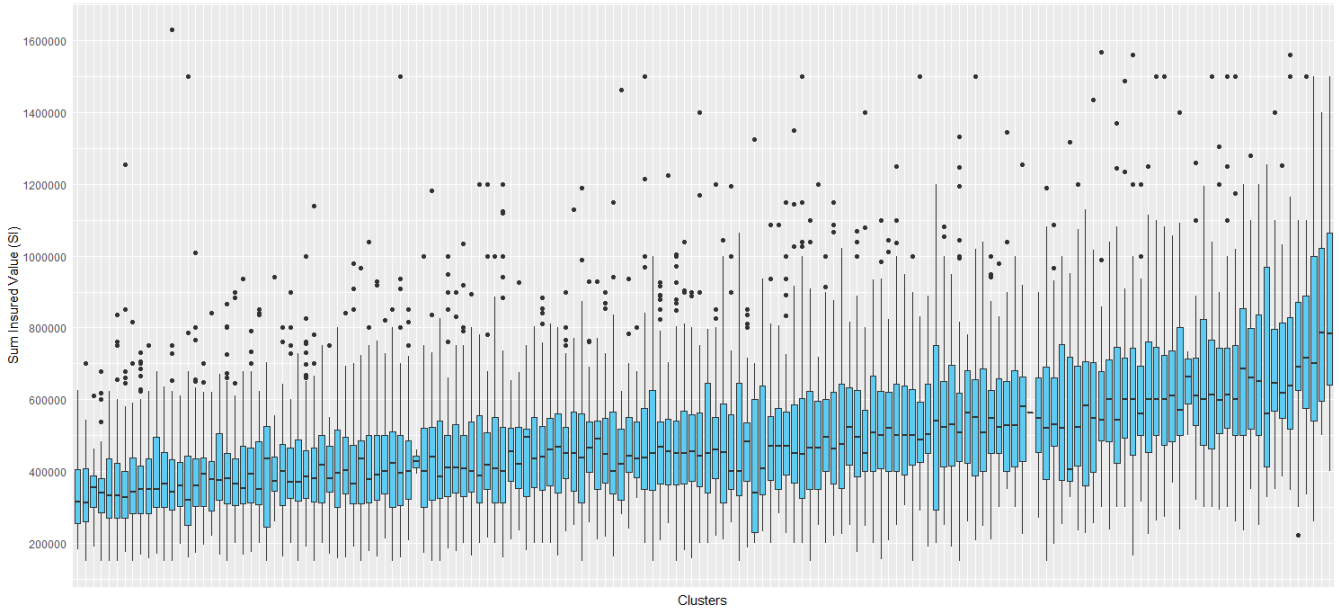

Fig. 4. Box plot of SI for each cluster. Ordered with respect to mean SI (including outliers).

To validate the choice of CLARA-PAM over other applications appendix **Table A.1** presents a summary criterion that captures similarity in medoids across iterations of the selected algorithms. As seen in **Table A.1**, PAM has the lowest mean and max distance from the medoid in building features out of the clustering algorithms tested. These results suggest there is limited additional value in devising a large application for the other algorithms (AGNES, H-Clust, k-means).

4.2 Predictive analysis

To condense the SIM information of each cluster to a single measure, we take the difference in the policyholder's SI value and the Huber-M adjusted mean SI of their cluster. A positive (negative) value for SIM refers to over-insurance (underinsurance); SIM reflects the dollar amount by which a customer's SI differs from the expected SI for their building. **Fig. 5** shows the relationship of the Huber-M mean SI for a given cluster. Computing SIM for each observation arrives at a portfolio median SIM of -$13,101.

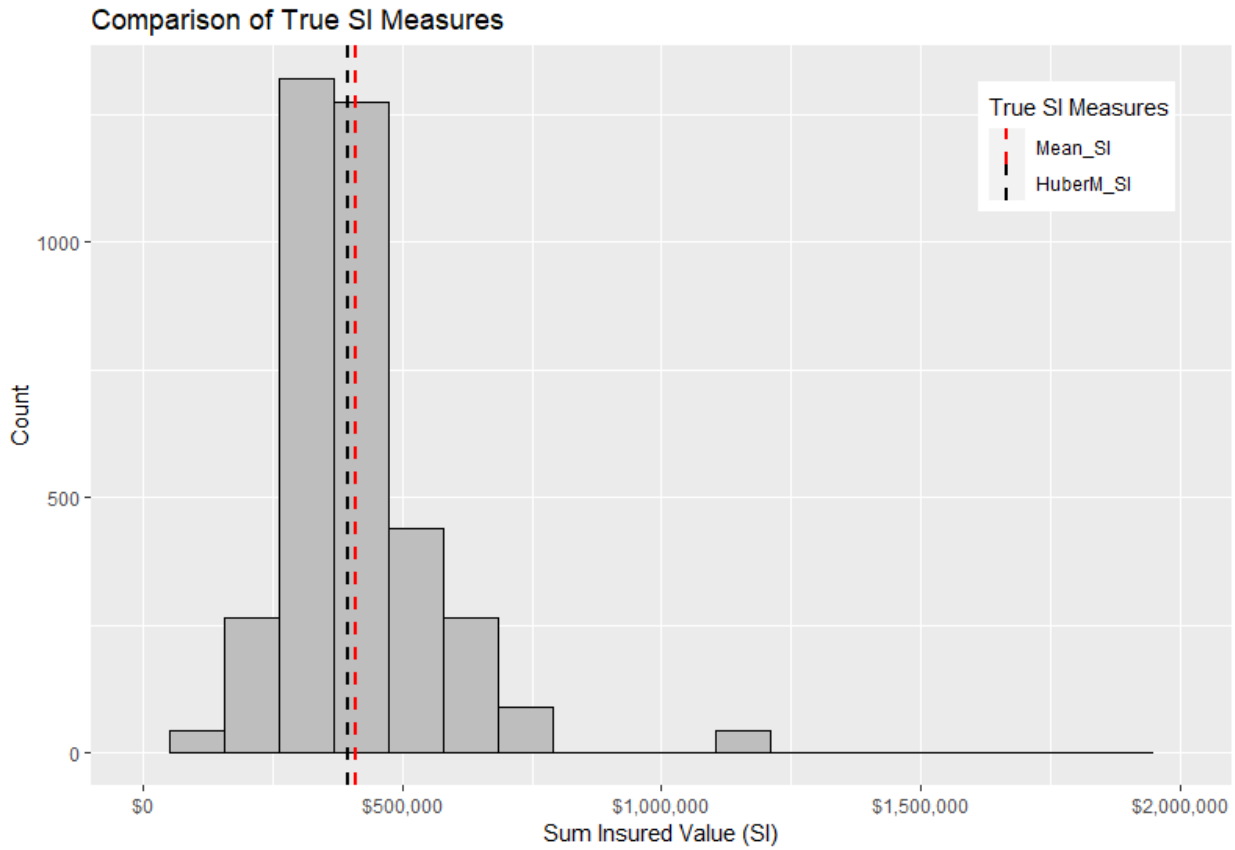

Fig. 5. An example of the SI distribution for cluster 3 shows that the Huber-M mean ignores the outlier with a SI of $1.2m. The difference between the Huber-M mean SI, and the building selected SI of $1.2m (distance from the black dotted line to the outlier), results in a larger SIM, than if we were to use the mean SI (distance from the red dotted line to the outlier). By employing the Huber-M as the true SI measure, the SIM for the outlier is larger, and thus we better capture SIM at the extremes.

The following sections then illustrate the multifaceted value of the proposed framework in a prediction perspective[4] through (i) understanding individual-level drivers or indicators of SIM (**Fig. 6**) and (ii) understanding portfolio-wide drivers of SIM (**Fig. 7** and **Fig. 8**):

- Waterfall chart: **Fig. 6** shows the directional impact of each feature on SIM prediction and how each feature contributes to the absolute value of SIM for an individual policyholder. Observing the impacts for individuals as opposed to population levels, is particularly crucial for the SIM problem. The consequences of underinsurance for a policyholder are severe, and by using SHAP values we observe the direct influence of certain factors for a specific policyholder. Stakeholders can devise a much more targeted response, where we can produce a bespoke impact chart for each policyholder across a national portfolio. Additionally, the results are easily interpretable as the SHAP value is in terms of the SIM, which is in dollar amounts. For example, given the current set of model features for the policyholder in **Fig. 6**, being in an area with a higher proportion of high school education attendance (high school education = 94th percentile) marginally contributes -$123,947 to the SIM prediction, and having a below median mortgage repayment (mortgage repayment = 46th percentile) marginally contributes +$113,282.
- SHAP importance summary: **Fig. 7** presents the SHAP value (x-axis) for each explanatory factor, ranked by importance in the model. The dots represent an individual policyholder, and they are shaded according to their feature value. This summary plot gives us a view of the most important variables driving SIM, but also leading indications into the relationships (as per the shaded feature values), and the materiality of potential outliers (with each dot representing an individual

[4] The prediction perspective refers to visualising the SHAP values produced from an XGBoost model that incorporated all socioeconomic and demographic features as listed in **Table 2**.

policyholder). For example, we can see that the two most important factors driving SIM are the policyholder age and drug rate information, which have a mean absolute impact of $31,952 and $29,667 respectively. We also see that a low drug rate (drug factor 2), where the blue dots indicate a low feature value, generally has a positive impact on SIM. The opposite occurs for a high drug rate. Outliers can also be identified, where having a low mortgage repayment, leads to a very high contribution to a positive SIM (over-insurance), with the impact over $200,000 for one policyholder. In this work, the average relationships are important to understand SIM on a national scale, but the outliers are the crucial focus given these are the policyholders that will be significantly over or underinsured.

- SHAP dependence plots: **Fig. 8** further illustrates the relationships summarised in **Fig. 7**, with the SHAP dependence plot shown individually for the four most important features in explaining SIM. The relationships are mostly intuitive, e.g. it is expected that younger policyholders are more likely to underinsure their building. The relationships observed are subject to a large amount of noise but provide a valuable insight to the average SIM experience. We can also contrast these findings with those in the literature [2,7,8,9]. To illustrate the value of the framework we provided the plots for the top four features only, but in practice it is useful to consider all features used in the predictive model.

In addition to the three primary charts, SHAP values also allow for contrasting explanations. Predictions can be compared to the average prediction of a subset of the data, or even a single data point, which is extremely useful when a particular factor is of interest. For example, insurers will be keen to understand whether their SIM Cover product tends to be used by policyholders that believe they may be underinsured, and whether the effects are different for a significant demographic factor such as age.

To summarise the value of the exploratory results of this framework we refer to the two core analysis pieces of the framework (Clustering and Predictive), and discuss the stakeholder-specific final aspect (Prescriptive):

- Clustering Analysis: provides a method to create meaningful unsupervised representations of policyholders insured buildings, from which we derived a metric for SIM. The clusters may also be useful in other actuarial functions such as rating factor selection for pricing.
- Predictive Analysis: provided inferential insights about SIM on a national scale and a framework to visualise predictive model results for the purpose of explaining SIM.
- Prescriptive Analysis: would serve to translate the predictive measure to meaningful business solutions. Various parties will use the information from prescriptive analysis for different means. For governments, the socioeconomic and demographic features that influence SIM is valuable in deciding policy action to minimise underinsurance. For insurance, prescriptive analysis drives the core processes of the business. The predictive outcomes can be used to facilitate more granular risk pricing and increase the effectiveness of product covers for policyholders that are likely to be underinsured.

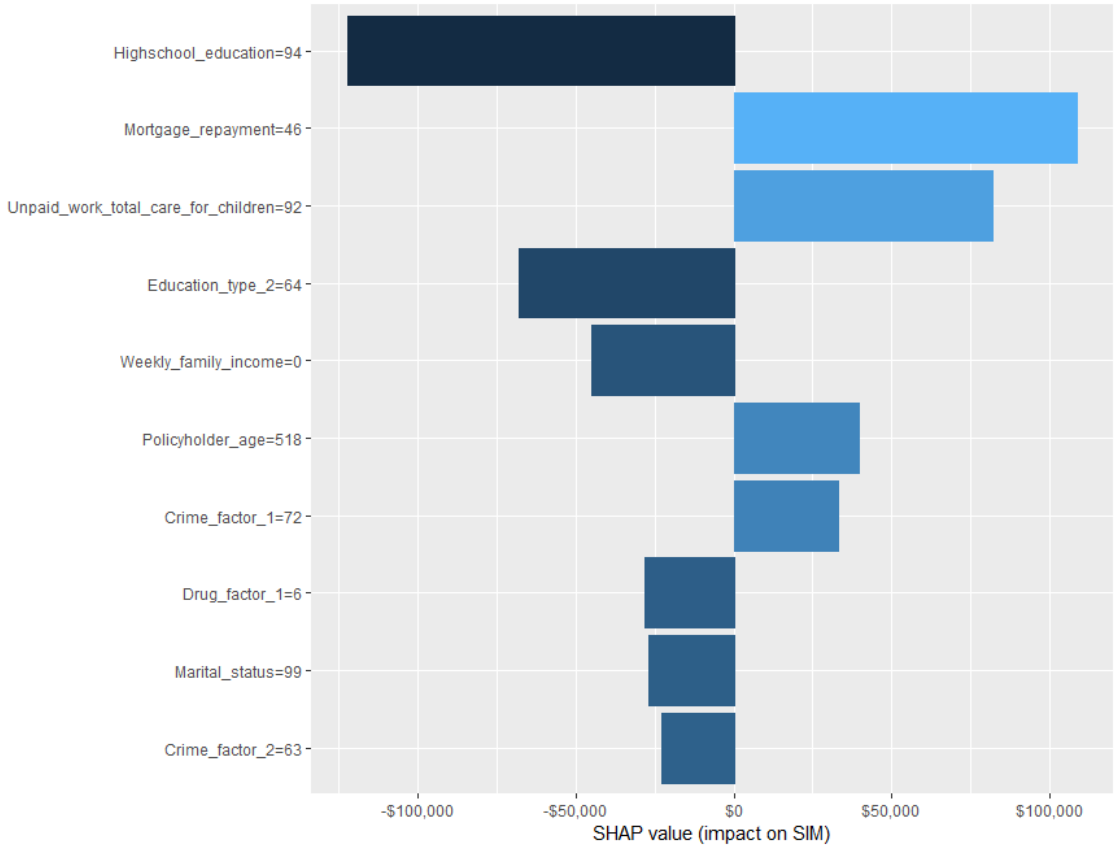

Fig. 6. Waterfall plot of variable contribution to SIM (SHAP values) for a policyholder with an SIM of -$60,722.

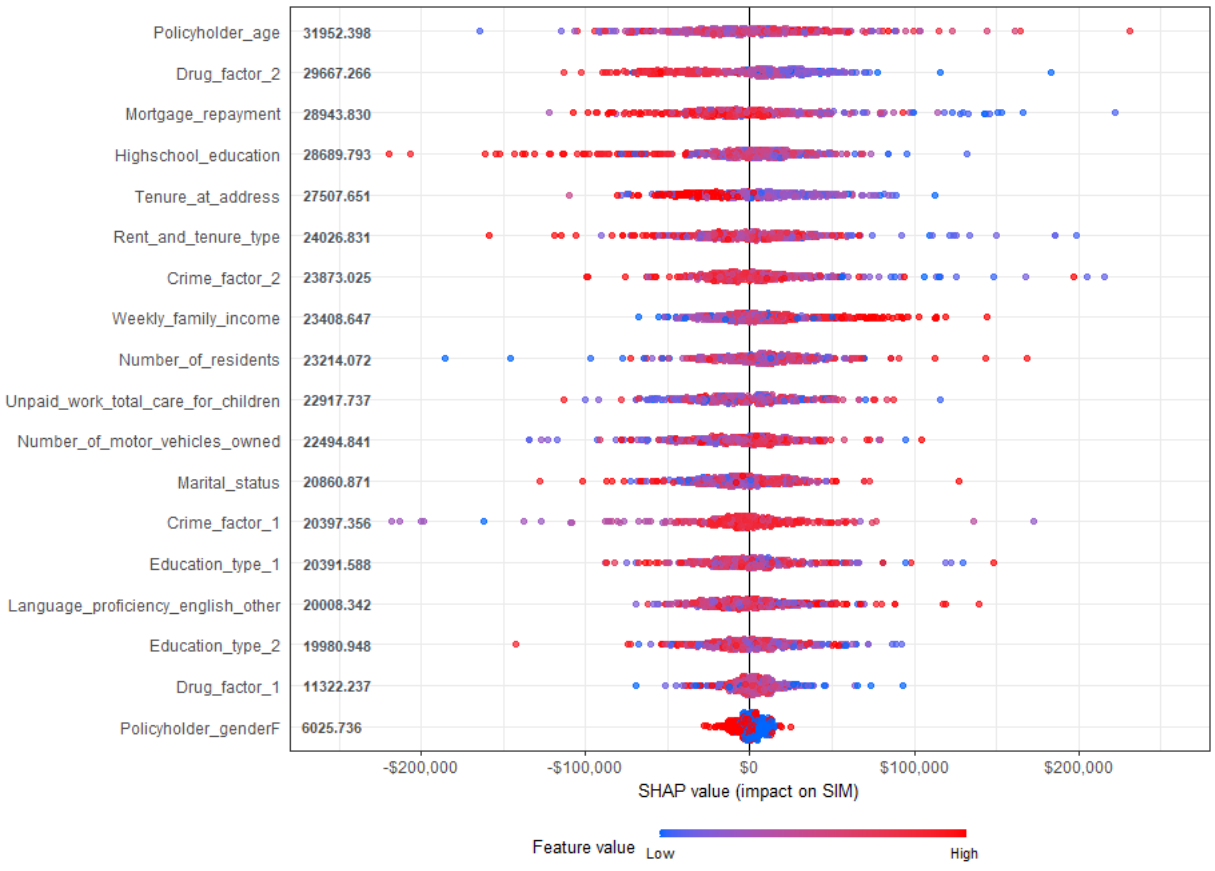

Fig. 7. Dependence of each attribute, showing the impact of each attribute from high to low on the SIM model prediction.

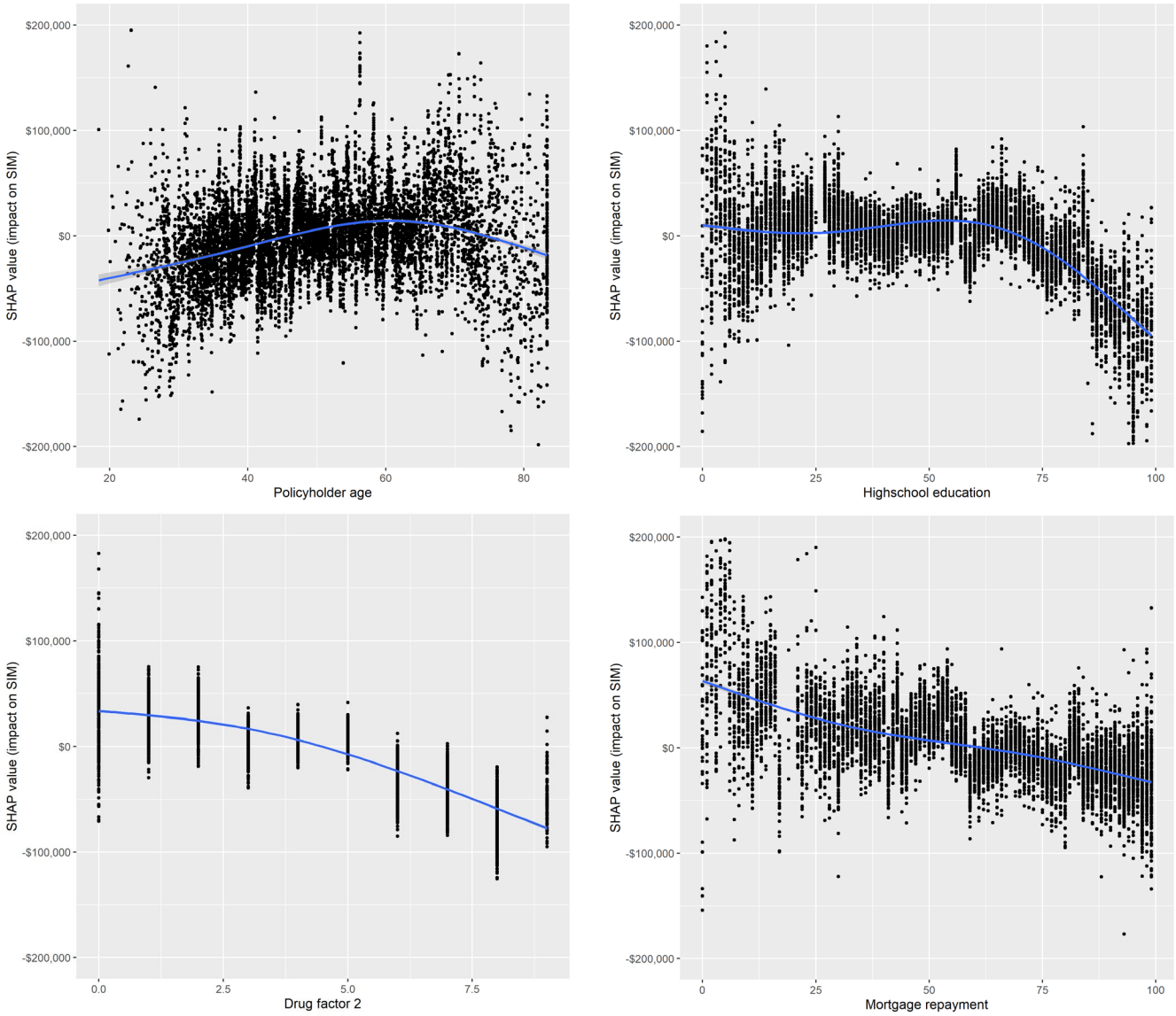

Fig. 8. Partial dependence plot for the four most important attributes in explaining SIM. Noting for Panel A – the Policyholder age was winsorized in the raw data (before the research process began) at 83.33 years (1000 months), this was expected to have limited impact on the results of the big data framework – where the distribution still holds 45 years (540 months) as the most common age. Noting for Panels B and D, the value is measured in percentile, while for Panel C it is decile.

5. Conclusion

Motivated by the ACCC's report on underinsurance [1], this paper provides an example of how data analytics can be used in insurance to further optimise and improve the customer experience, outside the realm of standard actuarial functions like price and retention optimisation. In fact, this is the first study in the literature to estimate individual SIM on a portfolio-wide scale.

The framework designed to investigate policyholder SIM provides valuable output to stakeholders at each stage of analysis. In this work, we used clustering analysis to diagnose SIM and predictive analysis to identify the socioeconomic and demographic factors that drive SIM. Complex relationships were carefully illustrated from analytical models in a manner that actions business decisions. It was found that the age of a policyholder, tenure at address, drug information (of the local area) and the mortgage information were important factors in explaining SIM. While not within the scope of this paper, the output from other stages of the framework could significantly improve insurance processes. For example, the clusters may be used as a rating factor in pricing models, with the potential to remove the

SI question altogether from the customers quote journey. The level of SIM may also be used in predictive models, ad-hoc analysis, and portfolio monitoring.

The framework employs a range of advanced data analytic techniques to harness large amounts of insurance data. We used FAMD to include a variety of data types, reduce dimensionality and avoid using computationally complex mixed data distance measures. CLARA-PAM was used to cluster an entire portfolio of policyholders, which streamlined and simplified future analyses by avoiding sampling approximations. To provide business insight, SHAP values illustrated the complex relationships from the XGBoost prediction model.

The framework has scope for many future studies of SIM characteristics. Future works will include greater breadth of building information to improve the accuracy of clusters, reference external sources of true SI to quantitatively validate SIM, and draw on the output from various stages of the framework for deeper insights on SIM.

Acknowledgements

The authors thank the Australian insurer (remaining anonymous) for providing the data for this research.

6. References

[1] Northern Australia Insurance Inquiry Final Report. 2020. Published by the Australian Competition & Consumer Commission under the Competition and Consumer Act 2010.

[2] Whittaker, J., Handmer, J. and Mercer, D., 2012. Vulnerability to bushfires in rural Australia: A case study from East Gippsland, Victoria. Journal of Rural Studies, 28(2), pp.161-173.

[3] Property Insurance Affordability: Challenges and Potential Solutions prepared by General Insurance Affordability Working Group on behalf of the Actuaries Institute Australia. Published November 2020.

[4] Cevolini, A., & Esposito, E. (2020). From pool to profile: Social consequences of algorithmic prediction in insurance. Big Data & Society, 7(2), 205395172093922.

[5] Ahmed, I., Ahmad, M., Jeon, G., & Piccialli, F. (2021). A Framework for Pandemic Prediction Using Big Data Analytics. Big Data Research, 25, 100190.

[6] Batarseh, F. A., & Latif, E. A. (2016). Assessing the Quality of Service Using Big Data Analytics. Big Data Research, 4, 13–24.

[7] Booth, K. and Tranter, B., 2017. When disaster strikes: Under-insurance in Australian households. Urban Studies, 55(14), pp.3135-3150.

[8] Hope, P. 2015. Market Adaptation to Climate Risk: Evaluating Property Insurance Pricing in Vulnerable Coastal Communities. University of Waterloo.

[9] Céline Grislain-Letrémy, 2018. Natural Disasters: Exposure and Underinsurance. Annals of Economics and Statistics, (129), p.53.

[10] Yeo, A. and Smith, K., 2003. Implementing a Data Mining Solution for an Automobile Insurance Company. Cases on Information Technology Series, pp.63-73.

[11] Samizadeh, R. and Mehregan, S. (2014). Retaining Customers Using Clustering and Association Rules in Insurance Industry: A Case Study. Int. J. Manag. Bus. Res., 5 (4), 261-268.

[12] Kašćelan, V., Kašćelan, L. and Novović Burić, M., 2016. A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market. Economic Research-Ekonomska Istraživanja, 29(1), pp.545-558.

[13] Williams, G. and Huang, Z., 1997. Mining the knowledge mine. Advanced Topics in Artificial Intelligence, pp.340-348.

[14]Khalili-Damghani, K., Abdi, F. and Abolmakarem, S., 2018. Solving customer insurance coverage recommendation problem using a two-stage clustering-classification model. International Journal of Management Science and Engineering Management, 14(1), pp.9-19.

[15] Dehghanpour, A. and Rezvani, Z., 2015. The profile of unethical insurance customers: a European perspective. International Journal of Bank Marketing, 33(3), pp.298-315.

[16] Carfora, M., Martinelli, F., Mercaldo, F., Nardone, V., Orlando, A., Santone, A. and Vaglini, G., 2018. A “pay-how-you-drive” car insurance approach through cluster analysis. Soft Computing, 23(9), pp.2863-2875.

[17] Devale, A., 2012. Applications of Data Mining Techniques in Life Insurance. International Journal of Data Mining & Knowledge Management Process, 2(4), pp.31-40.

[18] Zhuang, K., Wu, S. and Gao, X., 2018. Auto Insurance Business Analytics Approach for Customer Segmentation Using Multiple Mixed-Type Data Clustering Algorithms. Tehnicki vjesnik - Technical Gazette, 25(6).

[19] Yao, J. 2008. Clustering in Ratemaking: Applications in Territories Clustering. Discussion Paper Program. Casualty Actuarial Society.

[20] Khare, S., Bonazzi, A., Mitas, C. and Jewson, S., 2015. Modelling clustering of natural hazard phenomena and the effect on re/insurance loss perspectives. Natural Hazards and Earth System Sciences, 15(6), pp.1357-1370.

[21] Boriah, S., Chandola, V. and Kumar, V., 2008. Similarity Measures for Categorical Data: A Comparative Evaluation. Proceedings of the 2008 SIAM International Conference on Data Mining,.

[22]Chavent, M., Kuentz-Simonet, K., Labenne, A. and Saracco, J. 2017. Multivariate Analysis of Mixed Data: The R Package PCAmixdata. Computational Statistics.

[23] Schubert, E. and Rousseeuw, P. J. 2021. Fast and Eager k-Medoids Clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms. Information Systems, 101, p.101804.

[24] Menéndez, H. D., Otero, F. E. B., & Camacho, D. 2016. Medoid-based clustering using ant colony optimization. In Swarm Intelligence 10 (2) pp. 123–145. Springer Science and Business Media LLC. https://doi.org/10.1007/s11721-016-0122-5

7. Appendices

Appendix A Clustering Algorithm Comparisons

ID	Algorithm	Win rate	Mean Dist. from Medoid	Std. Dist. from Medoid	Max. Dist. from Medoid
1	K-means	0	1.390	0.052	1.477
2	PAM	100%	1.260	0.086	1.375
3	Agnes	0	1.863	0.150	2.102
4	H-Clust	0	1.517	0.072	1.669

Table A.1 Comparison of different clustering algorithms. We run each algorithm with a fixed random sample of 100 observations and randomly sample the remaining 9900 observations, across 20 iterations. The distance of the fixed observations to their medoid for each iteration are the basis for measures provided in the table. The win rate is the percentage of times out of 20, the algorithm had the lowest mean distance from medoid (across the fixed 100 observations).

Appendix B Multi-Criteria Evaluation of Number of Clusters

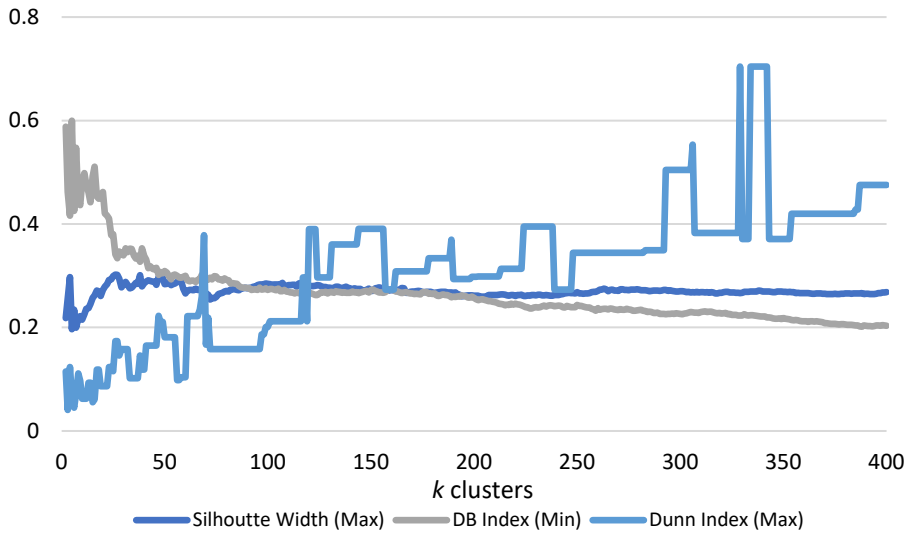

Fig. B.1 The grid range selected for k number of clusters was set as 2 to 400.